STA 302f20 Assignment Ten[1]

The following problems are not to be handed in. They are preparation for the quiz in tutorial and the final exam. Please try them before looking at the answers. Use the formula sheet. Be ready for R questions similar to the ones asked in this assignment. You will not be asked to hand in your complete answers to the R parts of this assignment, but you might be asked to do something similar on the quiz.

1. Based on the general linear model with normal error terms,
   (a) Prove the t distribution given on the formula sheet for a new observation y_0. Use earlier material on the formula sheet. For example, how do you know numerator and denominator are independent?
   (b) Derive the $(1-\alpha)\times 100\%$ prediction interval for a new observation from this population, in which the independent variable values are given in $\mathbf{x}_0$. "Derive" means show the High School algebra.

2. Suppose you have a random sample from a normal distribution, say $y_1, \ldots, y_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$. If someone randomly sampled another observation from this population and asked you to guess what it was, there is no doubt you would say $\overline{y}$, and a confidence interval for μ is routine. But what if you were asked for a *prediction* interval for the new observation?

   Accordingly, suppose the normal model is reasonable and you observe a sample mean of $\overline{y} = 7.5$ and a sample variance (with $n-1$ in the denominator) of $s^2 = 3.82$. The sample size is $n = 14$. Give a 95% prediction interval for the next observation. The answer is a pair of numbers. Show your work. You can get the distribution result you need from the formula sheet, or you can re-derive it for this special case. Do it both ways. You should use R to get the critical value.

3. A forestry company has developed a regression equation for predicting the amount of useable wood that they will get from a tree, based on a set of measurements that can be taken without cutting the tree down. They are convinced that a model with normal error terms is right. They have $\widehat{\boldsymbol{\beta}}$ and MSE based on a set of n trees they measured first and then cut down, and they know how to calculate a predicted y and a prediction interval for the amount of wood they will get from a single tree.

   But that's not what they want. They have a set of m more trees they are planning to cut down, and they have measured the predictor variables for each tree, yielding $\mathbf{x}_{n+1} \ldots, \mathbf{x}_{n+m}$. What they want is a prediction of the *total* amount of wood they will get from these trees, along with a 95% prediction interval for the total.
   (a) The quantity they want to predict is $w = \sum_{j=n+1}^{n+m} y_j$, where $y_j = \mathbf{x}_j'\boldsymbol{\beta} + \epsilon_j$. What is the distribution of w? You can just write down the answer without showing any work. The symbol w often stands for a chi-squared random variable, but here it means *wood*.
   (b) Let $\widehat{w}$ denote the prediction of w. It is calculated using the company's regression data along with $\mathbf{x}_{n+1} \ldots, \mathbf{x}_{n+m}$. Give a formula for $\widehat{w}$.
   (c) What is the distribution of $\widehat{w}$?
   (d) What is the distribution of $w - \widehat{w}$?
   (e) Now standardize $w - \widehat{w}$ to obtain a standard normal. Call it z.
   (f) Divide z by the square root of a chi-squared random variable, divided by its degrees of freedom, and simplify. Call it t. What are the degrees of freedom?
   (g) How do you know that numerator and denominator are independent?
   (h) Using your formula for t, derive the $(1-\alpha)\times 100\%$ prediction interval for u. Please use the symbol $t_{\alpha/2}$ for the critical value.

4. This question uses the built-in `trees` data set you saw in the lecture ("Regression Diagnostics with R"). We will continue to use the term Girth, even though we know that it's really diameter. Start by fitting a model with `Height`, `Girth` and `Girth` squared.

   The forestry company wants to predict the volume of wood they would obtain if they cut down three particular trees. The first tree has a girth of 11.0 and a height of 75. The second tree has a girth of 14.8 and a height of 80. The third tree has a girth of 10.5 and a height of 65. Using R,

   (a) Calculate a predicted amount of wood the company will obtain by cutting down these trees. The answer is a number.

   (b) Calculate a 95% prediction interval for the total amount of wood from the three trees. The answer is a pair of numbers, a lower prediction limit and an upper prediction limit. Take a look at `help(summary.lm)`. That `sigma` object is the square root of Mean Squared Error. It's called "residual standard error" in the output of `summary`.

5. For the general linear regression model, prove that the mean hat value equals $\frac{k+1}{n}$.

6. Consider a 'regression" model with an intercept, but no explanatory variables.

   (a) What is h_{ii}? They are all equal.

   (b) What is h_{ij}? They are all equal.

   Show some work.

7. Let $\mathbf{x}_i'$ denote row i of the $\mathbf{X}$ matrix. Essentially, it's the x data for person i. Show that $h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$. Why does this tell you that if the model has an intercept, all the hat values must be strictly greater than zero? Suggestion: Use $\mathbf{v}_i$ to denote an $n \times 1$ vector with all zeros except for a one in position i.

8. For a general multiple regression model with an intercept and k independent variables, show that the squared sample correlation between the y and $\widehat{y}$ values is equal to R^2. Thus, a scatterplot of $\widehat{y}_i$ versus y_i gives a picture of the strength of overall relationship between the independent variables and the dependent variable.

9. For the general linear regression model, are $\widehat{\mathbf{y}}$ and $\widehat{\boldsymbol{\epsilon}}$ independent?

   (a) Answer Yes or No and prove your answer.

   (b) What does this imply about the plot of predicted values against residuals?

10. For the general linear regression model, are $\mathbf{y}$ and $\widehat{\mathbf{y}}$ independent? Answer Yes or No and prove your answer.

11. For the general linear regression model, are $\mathbf{y}$ and $\widehat{\boldsymbol{\epsilon}}$ independent? Answer Yes or No and prove your answer.

12. A good question asked in lecture was why not plot the x values against $\widehat{y}$? For a simple regression (one explanatory variable) what would such a plot look like? Would it depend in any way on the correctness of the model?

13. Regression diagnostics are mostly based on the residuals, and transformations of the residuals. This question compares the error terms ϵ_i to the residuals $\widehat{\epsilon}_i$. Answer True or False to each statement. For statements about the residuals, show a calculation that proves your answer. You may use anything on the formula sheet.

    (a) $E(\epsilon_i) = 0$
    (b) $E(\widehat{\epsilon}_i) = 0$
    (c) $Var(\epsilon_i) = 0$
    (d) $Var(\widehat{\epsilon}_i) = 0$
    (e) ϵ_i has a normal distribution.
    (f) $\widehat{\epsilon}_i$ has a normal distribution.
    (g) $\epsilon_1, \ldots, \epsilon_n$ are independent.
    (h) $\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_n$ are independent.

14. One of these statements is true, and the other is false. Pick one, and show it is true with a quick calculation. Start with something from the formula sheet.

    - $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\epsilon}}$
    - $\mathbf{y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\epsilon}}$
    - $\widehat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \widehat{\boldsymbol{\epsilon}}$

    As the saying goes, "Data equals fit plus residual."

15. For the general linear regression model in which $\mathbf{X}$ is a matrix of constants, why does it not make sense to ask about independence of the predictor variable values and the residuals?

16. To diagnose problems with a regression model, it is standard practice to plot residuals against each explanatory variable in the model. Calculate the sample correlation between the x_{ij} values (for a general j) and the $\widehat{\epsilon}_i$ values, assuming $\sum_{i=1}^{n} \widehat{\epsilon}_i = 0$. This is why you would expect the scatterplot to show nothing but a shapeless cloud of points if the model is correct.

17. Show that if $\sum_{i=1}^{n} \widehat{\epsilon}_i = 0$, the sample correlation between the $\widehat{y}_i$ and $\widehat{\epsilon}_i$ values is exactly zero. Usually, when two random variables are independent, the distribution of the sample correlation is scattered around zero.

18. One of the built-in R datasets is `Puromycin`. Type the name to see it; there are only 23 lines of data. I believe the cases are test tubes; there were $n = 23$ test tubes. The dependent variable is the rate of a chemical reaction, specifically an enzymatic reaction. The test tubes contain cells and also a *substrate*, a reactant which is consumed during the enzymatic reaction. The independent variables are concentration of the substrate and whether or not the cells are treated with puromycin, an antibiotic.

    (a) Fit a model with just concentration and treatment with puromycin.
        i. Controlling for concentration of the substrate, does treatment with puromycin have an effect? If so, what is the effect on the rate of the chemical reaction? Of course you should be able to state the null hypothesis, and also give the numerical value of the test statistic and so on.
        ii. Controlling for treatment with puromycin, does concentration of the substrate affect the rate of the chemical reaction. If so, does higher concentration speed up the reaction, or slow it down?
        iii. You would not expect a straight-line relationship between concentration and rate of a chemical reaction. Verify this with a residual plot.

(b) There are better ways to analyze this data set, but let's do a rough version using polynomial regression. Do the right thing and fit another model.

   i. One of the default t-tests lets you verify that the relationship between concentration and rate is curvilinear. Which one? Give the value of the test statistic and the p-value.

   ii. How can you tell from the output of `summary` that the function is concave down?

(c) Plot the residuals from your second model against concentration. I think I see more curvyness, maybe even with two bends. I see a possible outlier too, but let it go for now. Add a cubic term to your regression model. The output of `summary` has a test for whether the cubic term significantly improves model fit. What do you conclude? Is the cubic term helpful?

(d) I think we should be fairly happy, because taken together, the polynomial terms improved the R^2 from 0.714 to 0.941. Plot the residuals again. Do you see that possible outlier?

(e) Treating the Studentized deleted residuals as test statstics and employing a Bonferroni correction, test for possible outliers.

   i. What is the Bonferroni critical value? This number should be on your printout. Be careful to get the degrees of freedom right.

   ii. Did you locate any outliers? For each one (if there are any), give the values of concentration and reaction rate (numbers), and state whether the cells were treated or untreated.